

**Diploma Universitario**

**en Ciencias Sociales Computacionales y Humanidades Digitales**

**Consigna Trabajo FInal**

**Opción 1 (Módulos 3 y 4)**

**Estudiante:** Carmen Tello

### Introducción

En este trabajo se tomará un subconjunto de las variables de la Encuesta Permanente de Hogares, con el objetivo de construir un modelo predictivo de la variable de Ingresos de la ocupación principal.

En el proceso intentaremos arrojar luz sobre el comportamiento de esta variable en relación con el resto de información contenida en la encuesta.

### Análisis exploratorio de los datos

El dataset cuenta con 15 variables:

* La variable objetivo, es decir la de Ingresos de la ocupación personal
* Dos variables identificadoras del hogar y vivienda, que dejaremos de lado ya que no aportan valor a nuestra modelización ni análisis.
* Seis variables sociodemográficas (sexo, edad, nivel educativo, aglomerado, relación con el jefe de hogar y estado civil)
* Seis variables laborales (sector del empleador, categoría ocupacional, intensidad ocupacional, horas trabajadas, rubro y calificación).

Dos de estas variables son numéricas (edad y horas trabajadas), mientras que las restantes son categóricas, algunas de ellas con ordinalidad.

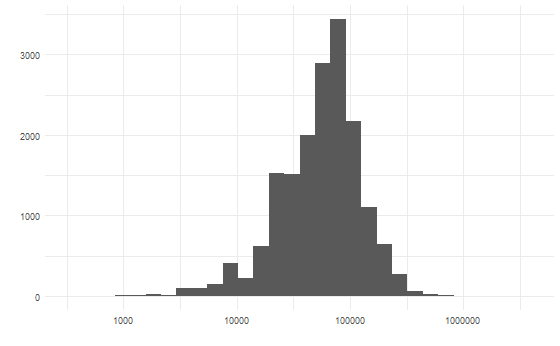
Haremos una breve descripción de la composición de cada variable y su relación con la variable objetivo.

Ingreso

En primer lugar, hay 3564 casos que registran un valor de -9. Estos casos se corresponden con personas que eligieron no responder esta pregunta. En general, quienes no responden esta pregunta tienden a tener ingresos muy bajos o muy altos. Excluiremos estos casos del análisis ya que no aportarán valor sino ruido en nuestra modelización.

Esto nos deja con un dataset de 17800 casos para trabajar. En este subgrupo, es una variable que tiene como valor mínimo 0 y máximo 3500000. El valor más típico es 60.000, y la media es de 70.695.

**Gráfico 1. Histograma de la variable Ingreso**

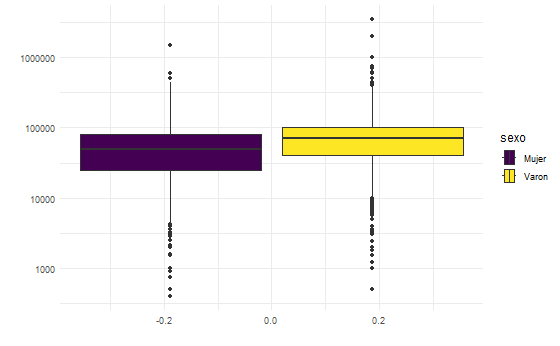


Además, vemos una distribución de la variable más orientada hacia la izquierda, es decir que los ingresos más bajos son más frecuentes que los altos.

Sexo

Al observar la distribución de la variable sexo, vemos que hay más hombres (54,9%) que mujeres (45,1%) en el dataset. Esto se corresponde con las teorías que establecen que en general los hombres tienen mayor inserción en el mercado laboral.

Además, al modelar una Regresión Lineal entre la variable de sexo y la de ingreso vemos que hay significancia estadística, y el estimate positivo para la categoría Varón demuestra sugiere que los varones tienden a cobrar más que las mujeres.

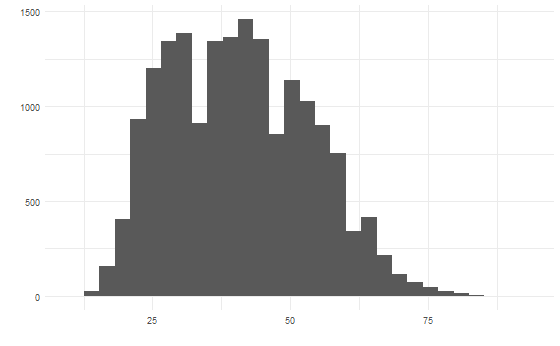


Esto se puede ver en el boxplot, donde la distribución de los varones es en general más alta que la de las mujeres.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 60697.33 | 737.1868 | 82.33643 | 0.0000000000000000000000000000000000000000000000000000000000000000000000000000000 |
| sexoVaron | 18215.49 | 995.0393 | 18.30630 | 0.0000000000000000000000000000000000000000000000000000000000000000000000000353346 |

Edad

Vemos una distribución relativamente uniforme, con un quiebre a partir de los 60/65 años, lo que se corresponde con la edad jubilatoria.

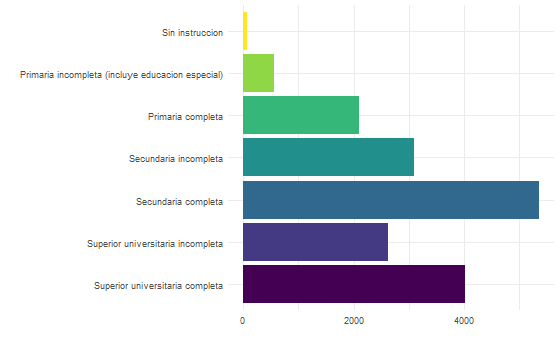


Al correr una regresión lineal, vemos una significancia estadística muy alta en la relación entre la edad y el ingreso.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 38993.6902 | 1644.35863 | 23.71362 | 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000002054256 |
| edad | 778.0842 | 38.49366 | 20.21331 | 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000761459810689933786452543440503859528690 |

Nivel educativo

Se observa una correlación entre el nivel educativo y el ingreso, de manera que a mayor nivel educativo crece el ingreso. Dado que la variable tiene 7 categorías y una distribución poco equitativa (por ejemplo, las categorías de Primaria incompleta y sin instrucción tienen un 3% y 0.3% de frecuencia respectivamente.), construiremos una nueva variable, donde agruparemos las categorías en 3: Alto (estudios universitarios incompletos o más), Medio (estudios secundarios completos o incompletos) y Bajo (estudios primarios completos o menos) para reducir el ruido que pueda aportar al modelo.



| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 104949.36 | 1003.373 | 104.596543 | 0.000000e+00 |
| nedSuperior universitaria incompleta | -36409.29 | 1598.009 | -22.784156 | 2.740423e-113 |
| nedSecundaria completa | -37470.30 | 1327.568 | -28.224768 | 1.711320e-171 |
| nedSecundaria incompleta | -52464.86 | 1521.841 | -34.474611 | 3.569593e-252 |
| nedPrimaria completa | -54207.62 | 1714.100 | -31.624542 | 1.335961e-213 |
| nedPrimaria incompleta | -61962.87 | 2883.388 | -21.489601 | 3.752717e-101 |
| nedSin instruccion | -56214.44 | 8078.581 | -6.958455 | 3.560032e-12 |

Al recategorizar, la correlación se mantiene: vemos un p valor que muestra una significancia estadística alta, y vemos con los valores del `estimate` que a menor nivel educativo el ingreso disminuirá.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 90595.17 | 792.3775 | 114.33335 | 0.000000e+00 |
| ned\_catMedio | -23116.11 | 1185.6907 | -19.49590 | 8.890331e-84 |
| ned\_catBajo | -39687.06 | 1160.0950 | -34.21018 | 1.789620e-248 |

Aglomerado

Aglomerado es una variable que tiene 32 categorías.

Algunos aglomerados tienen significancia estadística, mientras que otros no.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 82131.9242 | 3521.299 | 23.32432610 | 1.509923e-118 |
| aglomeradoUshuaia - Rio Grande | 51056.4816 | 4720.513 | 10.81587583 | 3.518421e-27 |
| aglomeradoCdro. Rivadavia - R.Tilly | 41807.6970 | 4852.731 | 8.61529186 | 7.552651e-18 |
| aglomeradoS. del Estero - La Banda | -34267.9672 | 4321.503 | -7.92964059 | 2.327315e-15 |
| aglomeradoGran Tucuman - T. Viejo | -32276.2726 | 4153.036 | -7.77172914 | 8.162762e-15 |
| aglomeradoLa Rioja | -31778.2466 | 4336.250 | -7.32850821 | 2.427115e-13 |
| aglomeradoCiudad de Buenos Aires | 31000.2197 | 4290.470 | 7.22536608 | 5.200074e-13 |
| aglomeradoGran Resistencia | -32873.4880 | 4562.783 | -7.20470079 | 6.050211e-13 |
| aglomeradoJujuy - Palpala | -26141.2700 | 4226.421 | -6.18520327 | 6.337848e-10 |
| aglomeradoGran Catamarca | -27325.8883 | 4445.108 | -6.14740647 | 8.044093e-10 |
| aglomeradoCorrientes | -26280.4985 | 4515.688 | -5.81982117 | 5.992436e-09 |
| aglomeradoGran Mendoza | -22701.2324 | 4188.858 | -5.41943292 | 6.056698e-08 |
| aglomeradoFormosa | -24709.0327 | 4717.990 | -5.23719522 | 1.648937e-07 |
| aglomeradoSalta | -19830.5855 | 4098.705 | -4.83825603 | 1.320818e-06 |
| aglomeradoConcordia | -22167.5508 | 4738.484 | -4.67819456 | 2.915437e-06 |
| aglomeradoPosadas | -18262.5365 | 4555.245 | -4.00912242 | 6.119371e-05 |
| aglomeradoGran Cordoba | -14122.3059 | 4106.307 | -3.43917411 | 5.848308e-04 |
| aglomeradoRio Gallegos | 17464.4394 | 5222.930 | 3.34380094 | 8.280982e-04 |
| aglomeradoSanta Rosa - Toay | -16618.1422 | 5051.183 | -3.28995048 | 1.004000e-03 |
| aglomeradoGran San Juan | -12669.1983 | 4263.683 | -2.97142102 | 2.968219e-03 |
| aglomeradoSan Luis - El Chorrillo | -12836.1522 | 4463.332 | -2.87591269 | 4.033401e-03 |
| aglomeradoPartidos del GBA | -10230.5728 | 3852.400 | -2.65563632 | 7.922956e-03 |
| aglomeradoMar del Plata - Batan | -11125.0689 | 5026.629 | -2.21322656 | 2.689469e-02 |
| aglomeradoRio Cuarto | -8106.6990 | 4649.239 | -1.74366143 | 8.123544e-02 |
| aglomeradoGran Parana | -8578.1074 | 5293.137 | -1.62060950 | 1.051192e-01 |
| aglomeradoNeuquen - Plottier | 6862.0698 | 4968.640 | 1.38107603 | 1.672730e-01 |
| aglomeradoSan Nicolas - Villa Constitucion | -5991.0093 | 4594.084 | -1.30407041 | 1.922265e-01 |
| aglomeradoGran Santa Fe | -6074.8910 | 4781.698 | -1.27044639 | 2.039423e-01 |
| aglomeradoGran La Plata | -5189.1185 | 4712.975 | -1.10102815 | 2.708993e-01 |
| aglomeradoGran Rosario | -4323.5909 | 4416.271 | -0.97901387 | 3.275865e-01 |
| aglomeradoRawson - Trelew | 2443.7337 | 4623.084 | 0.52859387 | 5.970938e-01 |
| aglomeradoViedma - Carmen de Patagones. | -207.9611 | 4672.020 | -0.04451203 | 9.644968e-01 |

32 rows

Veamos qué sucede al agrupar esta variable según región. Tomaremos las regiones construidas por el [Ministerio de Economía](https://www.argentina.gob.ar/economia/politicaeconomica/regionalysectorial/economiasregionales/icer), pero separando el AMBA del resto de Buenos Aires.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 83814.024 | 1338.824 | 62.602741 | 0.000000e+00 |
| regionNOA | -29852.726 | 1661.091 | -17.971762 | 1.392427e-71 |
| regionNEA | -27239.127 | 2003.128 | -13.598294 | 6.639309e-42 |
| regionCuyo | -18291.840 | 1963.758 | -9.314710 | 1.362130e-20 |
| regionPatagonia | 13527.404 | 1850.140 | 7.311559 | 2.752673e-13 |
| regionCentro | -11921.682 | 1735.230 | -6.870375 | 6.615294e-12 |
| regionMar del Plata | -12807.169 | 3883.038 | -3.298234 | 9.748590e-04 |
| regionResto Buenos Aires | -4578.837 | 2729.128 | -1.677766 | 9.341045e-02 |

La relación se mantiene y logramos reducir las categorías a 7.

Estado civil

Una variable con una distribución poco equitativa, donde la categoría “Viudo/a” es marginal (representa un 1,3% del dataset), pero muestra una relación con la variable de ingresos.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 61052.439 | 3625.779 | 16.838435 | 3.931057e-63 |
| estado\_civilseparado/a o divorsiado/a? | 9569.866 | 3971.842 | 2.409428 | 1.598764e-02 |
| estado\_civilcasado? | 25992.616 | 3754.415 | 6.923214 | 4.565734e-12 |
| estado\_civilunido? | 11564.489 | 3733.447 | 3.097537 | 1.954397e-03 |
| estado\_civilsoltero/a? | -4282.265 | 3726.072 | -1.149270 | 2.504599e-01 |

Relación con el jefe de hogar

Esta variable tiene 10 categorías con distribuciones muy poco equitativas, y no todas muestran una relación fuerte con la variable de ingreso. Recategorizaremos esta variable en cuatro categorías: "Jefe/a", "Conyuge / Pareja", "Hijo/a Hijastro/a", "Otro".

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 73454.32 | 1017.954 | 72.158759 | 0.000000e+00 |
| relacion\_jefe\_catHijo/a Hijastro/a | -26999.83 | 1489.234 | -18.130017 | 8.348375e-73 |
| relacion\_jefe\_catJefe/a | 7937.11 | 1227.283 | 6.467221 | 1.024198e-10 |
| relacion\_jefe\_catOtro | -22093.48 | 2231.175 | -9.902172 | 4.672811e-23 |

Calificación

Lo mismo sucede con la variable de calificacion, que tiene algunas categorías poco comunes y aparentemente poco explicativas que expresan falta de información. Recategorizaremos en "Profesionales", "Técnicos", "Operativos", "No calificados", y "Otro".

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 139169.36 | 1724.771 | 80.688622 | 0.000000e+00 |
| calificacionTécnicos | -51223.65 | 2047.538 | -25.017186 | 8.752430e-136 |
| calificacionOperativos | -70421.83 | 1837.278 | -38.329430 | 6.322326e-309 |
| calificacionNo calificados | -101263.32 | 1997.463 | -50.695958 | 0.000000e+00 |
| calificacionOtro | -69893.92 | 8372.052 | -8.348482 | 7.416034e-17 |

5 rows

Categoría ocupacional

Esta variable parece estar muy relacionada con el ingreso.

| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 117117.02 | 2852.138 | 41.06289 | 0.000000e+00 |
| cat\_ocupacionalCuenta propia | -67071.72 | 3047.617 | -22.00793 | 6.238755e-106 |
| cat\_ocupacionalObrero o empleado | -42088.61 | 2906.880 | -14.47896 | 3.054280e-47 |
| cat\_ocupacionalTrabajador familiar sin remuneracion | -117117.02 | 7340.394 | -15.95514 | 6.511055e-57 |

4 rows

Sector empleador

Dentro de la variable sector\_empleador, la categoría "...privada", que representa el 73,8% del dataset, no tiene significancia estadística en correlación con la variable de ingreso. Esta variable podría no aportar un gran valor.

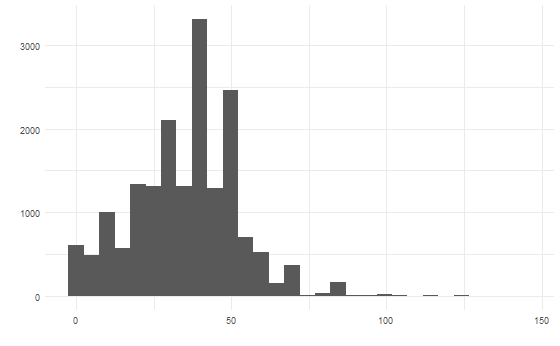
| **term**  <chr> | **estimate**  <dbl> | **std.error**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| (Intercept) | 64875.8203 | 4103.103 | 15.8114026 | 6.239130e-56 |
| sector\_empleador...estatal | 26192.0704 | 4220.726 | 6.2055838 | 5.569920e-10 |
| sector\_empleador...privada | -890.9887 | 4142.874 | -0.2150654 | 8.297188e-01 |

3 rows

Horas trabajadas

Vemos que la variable `horas\_trabajadas` tiene un valor de "999", lo que corresponde a una no-respuesta. Vamos a reemplazar este valor por el más común de la variable, que es 40.

Los casos se vuelven atípicos a partir de las 75 horas semanales.



Rubro

Una vez hecho este análisis preliminar, un paso más para poder elegir las variables predictoras es chequear la multicolinealidad, para seguir reduciendo el ruido que puedan aportar las variables a nuestro modelado.

Vemos que la variable de rubro muestra un fuerte grado de colinealidad, de modo que la quitamos de un primer modelo y volvemos a chequear. Esta estracción también corrige los valores altos que mostraba la variable cat\_ocupacional y calificacion, ya que todas las variables muestran valores menores a 3.

GVIF Df GVIF^(1/(2\*Df))

sexo 1.277751 1 1.130377

edad 1.611647 1 1.269507

estado\_civil 2.353110 4 1.112898

sector\_empleador 1.385026 2 1.084837

cat\_ocupacional 1.553298 3 1.076157

int\_ocupacional 2.825931 3 1.189032

horas\_trabajadas 2.829674 1 1.682163

calificacion 2.000573 4 1.090547

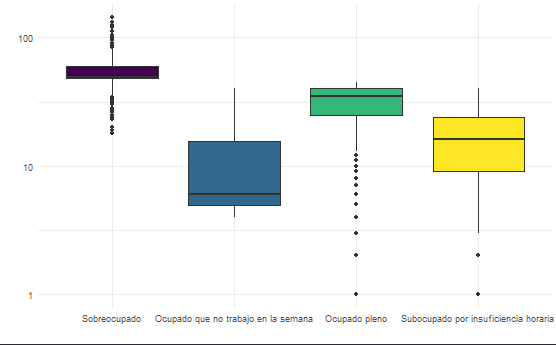
ned\_cat 1.567551 2 1.118937

region 1.105230 7 1.007172

relacion\_jefe\_cat 2.274902 3 1.146816

Intensidad ocupacional

También tengo la intuición de que hay dos variables que refieren a la misma información: la de intensidad ocupacional y la de horas trabajadas. Al establecer una regresión lineal, vemos que hay significancia estadística entre todas las categorías. Será prudente probar un modelo que no incluya la variable de intensidad ocupacional.



### Preparación final de la base

Con el objetivo de comenzar el proceso de modelización, vamos a dividir nuestra base trabajada en dos: un dataset de entrenamiento con el que construiremos el modelo, y un dataset de test con el que analizaremos la performance del modelo. A su vez, la base de entrenamiento va a ser dividida en 10 *folds*, para evitar los sesgos que puede contener la muestra al ser seleccionada.

### Modelización

Una vez hecho esto, se está en condiciones de comenzar el modelado de los datos. Elegimos trabajar con regresión lineal y con random forest, con el objetivo de entender si el vínculo que hay entre el ingreso y el resto de las variables responde más a un movimiento lineal (donde al aumentarse en una unidad una variable, el target también aumenta una unidad y viceversa), o si por el contrario responde a diferentes combinaciones de variables.

Regresión lineal

Vamos a probar tres recetas que tienen distintas combinaciones de variables:

* predictores\_originales: con todas las variables disponibles previas a las transformaciones del EDA.
* predictores\_originales2: con las variables originales y excluyendo las variables de `rubro`, de `int\_ocupacional` (por la presunta correlación con la variable de horas\_trabajadas), y de `sector\_empleador` y `estado\_civil` (ya que no mostraron mucha relación con la variable de ingresos)
* predictores\_transformados: con las variables transformadas en el EDA, y excluyendo la variable de `rubro` que mostró una fuerte colinealidad.
* Una con las variables transformadas en el EDA y excluyendo las variables de `rubro`, de `int\_ocupacional`, de `sector\_empleador` y de `estado\_civil`.

Además, todas las recetas incluyen los siguientes steps:

* step\_dummy: para convertir todas las variables categóricas en dummies, es decir en variables numéricas que facilitan el procesamiento.
* step\_normalize: para normalizar todas las variables numéricas, y evitar el ruido que puede generar que variables numéricas tengan escalas muy disímiles
* step\_zv: para excluir del modelado aquellas variables que tienen frecuencias muy altas del valor 0 o nulos. Esto es particularmente importante de realizar luego de convertir las variables a dummies, ya que pueden emerger categorías muy poco frecuentes como variables que no aportan demasiado valor al modelado.

Dentro de estas recetas, seleccionaremos la que corresponde a predictores\_originales, ya que presenta los mejores valores de raíz del error cuadrático medio (rmse) y de R cuadrado (sqr). Aplicaremos este modelo a nuestro dataset de test, con el objetivo de poder comparar los resultados con el siguiente modelo.

Random forest

Vamos a probar una receta que incluye todas las variables del dataset, incluídas las transformadas en el preprocesamiento del EDA.

Además de los steps ya mencionados, aplicaremos el step\_other que construye una categoría “otro” a partir de un cierto threshold de frecuencia. En este caso, las categorías que tengan una frecuencia menor al 1% serán transformadas en un “Otro”.

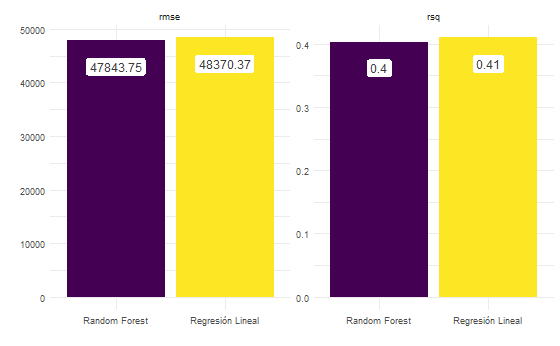
Por último, tuneamos los hiperparámetros mtry (la cantidad de predictores que se tomarán en cada división del árbol), y min\_n (número de data points requeridos por el nodo para dividir a continuación). Es decir, las dos çaracterísticas fundamentales del random forest, la aleatoriedad de los predictores y de los splits esperados.

Seleccionaremos el árbol que mejor performa según la métrica de rmse.

### Conclusiones

Testeando los modelos

Al comparar la performance de cada modelo aplicado sobre el dataset de test, podemos ver que los resultados son muy similares, apenas en favor de la regresión lineal.



Nos resulta esperable que pueda haber una relación lineal entre las variables trabajadas y el target, ya que el ingreso está muy atravesado por la calificación (nivel educativo, categoría ocupacional, calificación, e incluso edad), el género (un mercado laboral patriarcal que tiende a dar remuneraciones más bajas a las mujeres), el lugar de residencia (en cuanto a calidad de vida, condiciones habitacionales o incluso el costo de vida), etc.

En este sentido, creemos que no es necesario utilizar modelos más complejos para resolver este problema: el Random Forest requiere de mayor poder de cómputo, mayor tiempo y mayor complejidad analítica para ser construído, y no está aportando un diferencial en la capacidad predictiva.

Sin embargo, lo más interesante que tiene este modelo para aportar está vinculado a la Feature Importance. Este modelo permite arrojar un poco de luz acerca de las variables que están más involucradas en la predicción, lo que significa entender qué variables explican más o mejor a nuestro target.  
En este caso, podemos ver que algunas de las variables más importantes son: la edad, las horas trabajadas y la no-calificación.